Ultrasound in Med. & Biol., Vol. 42, No. 1, pp. 51–59, 2016

0301-5629/$ - see front matter

http://dx.doi.org/10.1016/j.ultrasmedbio.2015.08.005

● *Original Contribution*

VARIABILITY IN INTERPRETATION OF ULTRASOUND ELASTOGRAPHY AND GRAY-SCALE ULTRASOUND IN ASSESSING THYROID NODULES

Jieun Koh,* Hee Jung Moon,* Jeong Seon Park,† Soo Jin Kim,‡ Ha Yan Kim,§ Eun-Kyung Kim,* and Jin Young Kwak*

*Department of Radiology, Research Institute of Radiological Science, College of Medicine, Yonsei University, Seoul, Korea; †Department of Radiology, College of Medicine, Hanyang University, Seoul, Korea; ‡Department of Radiology, College of Medicine, Chung-Ang University, Seoul, Korea; and §Biostatistics Collaboration Unit, College of Medicine, Yonsei University, Seoul, Korea

(*Received* 17 *July* 2014; *revised* 5 *August* 2015; *in final form* 7 *August* 2015)

Abstract—The aim of this study was to validate inter-observer variability for strain ultrasound elastography (USE) and to compare the diagnostic performance of a combination of gray-scale ultrasound (US) and USE with that of gray-scale US. Three observers from different institutions evaluated gray-scale US images and USE video files of 443 cytopathologically proven benign or malignant thyroid nodules over a 3-mo period. Inter-observer variability did not statistically differ between USE using the Asteria criteria and gray-scale US; however, USE using the Rago criteria had the lowest inter-observer agreement ($p < 0.043$). For all three observers, sensitivity was increased by adding USE to gray-scale US (81.3%–88.3%, 75.4%–85.4%) compared with gray-scale US (70.4%–80.8%). Specificity was decreased by adding USE to gray-scale US (51.7%–59.1%, 59.1%–73.9%) compared with gray-scale US (69.0%–82.8%). USE and gray-scale US had comparable inter-observer variability. However, on addition of USE to gray-scale US, the additional diagnostic yield was limited compared with that of gray-scale US alone. (E-mail: docjin@yuhs.ac) © **2016 World Federation for Ultrasound in Medicine & Biology.**

Key Words: **Elastography, Thyroid nodule, Inter-observer variability, Ultrasound, Diagnosis.**

INTRODUCTION

Gray-scale ultrasound (US) is the most sensitive test currently available for detecting thyroid lesions; however, differentiation of benign and malignant nodules is not highly accurate with gray-scale ultrasound (Takashima et al. 1995), and thus, its diagnostic value varies considerably from study to study (Fish et al. 2008; Frates et al. 2006; Kim et al. 2002; Kovacevic and Skurla 2007; Lim et al. 2012). Ultrasound elastography (USE) enables the assessment of tissue consistency by differentiating stiff nodules from soft nodules, and it supplements the diagnostic limitations of gray-scale US (Asteria et al. 2008; Azizi et al. 2013; Bamber et al. 2013; Cantisani et al. 2014; Hong et al. 2009; Kagoya et al. 2010; Kim et al. 2014; Mehrotra et al. 2013; Moon et al. 2012; Rago et al. 2007; Rubaltelli et al. 2009; Shuzhen 2012; Shweel and Mansour 2013; Trimboli et al. 2012; Unluturk et al. 2012; Yoon et al. 2014). Previous studies suggested that with respect to diagnostic performance, USE is better or comparable to gray-scale US when differentiating benign from malignant thyroid nodules (Asteria et al. 2008; Azizi et al. 2013; Cantisani et al. 2014; Hong et al. 2009; Rago et al. 2007; Shuzhen 2012; Trimboli et al. 2012). Diagnostic performance was also improved with the combination of gray-scale US and USE (Shweel and Mansour 2013). Contrary to these positive results, several studies have failed to prove the superiority of USE to gray-scale US (Kagoya et al. 2010; Ko et al. 2014; Moon et al. 2012; Unluturk et al. 2012). Moreover, the combination of USE and gray-scale US was found to be inferior to gray-scale US in certain cases (Moon et al. 2012).

In addition to variable diagnostic performance, another technical issue with USE is limited inter-observer agreement (Kwak and Kim 2014), a problem first reported by Park et al. (2009) for strain USE. However, that study did not use subjective methods for monitoring compression. Other consecutive studies reported

Address correspondence to: Jin Young Kwak, Department of Radiology, Research Institute of Radiological Science, College of Medicine, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, 120-752 Seoul, Korea. E-mail: docjin@yuhs.ac

increased inter-observer agreement using subjective methods for monitoring compression on strain USE (Calvete et al. 2013; Merino et al. 2011; Ragazzoni et al. 2012). Shear wave USE has been reported to have fair to excellent reproducibility in neck lesions, including thyroid nodules, with higher inter-observer agreements compared with strain USE (Bhatia et al. 2012; Veyrieres et al. 2012; Zhang et al. 2012). To date, studies evaluating the inter-observer variability of strain USE have been limited by small sample size and have been performed only by observers from the same institution (Calvete et al. 2013; Merino et al. 2011; Park et al. 2009; Ragazzoni et al. 2012). Therefore, we focused on validating the inter-observer agreement for strain USE as well as gray-scale US by having three radiologists from different institutions compare the diagnostic performance of gray-scale US and a combination of gray-scale US and USE in a relatively large number of thyroid nodules.

METHODS

Patients

The institutional review board approved this retrospective study and required neither patient approval nor informed consent for our review of patient images and records. From November 2011 to January 2012, 583 nodules in 465 consecutive patients underwent fine-needle aspiration (FNA) or staging US with strain USE. We excluded nodules measuring <5 mm or ≥ 30 mm (n = 65) and nodules for which cytology results were classified as suspicious malignant (n = 17), atypia (n = 32), follicular neoplasm (n = 2) and non-diagnostic results (n = 10) with no further surgical intervention. Among 457 nodules, 194 were pathologically confirmed by surgery, and 263 were cytologically proved to be benign or malignant with no further surgical intervention. Among them, 14 nodules were excluded because of the poor quality of USE video files or gray-scale US images. Finally, 443 nodules in 426 patients were included in this study; among these 426 patients, 17 patients had two nodules. Mean age was 47 ± 12 y; 347 patients were female, and 79 were male. The mean size of the nodules was 11 ± 5.6 mm; 212 nodules were ≤ 10 mm, and the remainder were >10 mm.

Gray-scale US and USE

Gray-scale US was performed initially with a 6- to 14-MHz linear array transducer (EUB-7500, Hitachi Medical, Tokyo, Japan) by seven radiologists with 1 to 15 y of experience in thyroid imaging. Transverse and longitudinal images of thyroid nodules were captured and stored for subsequent image interpretation. After gray-scale US, USE was performed by the same radiologist with the same US unit. All USE images were obtained in longitudinal planes with the freehand technique. Each radiologist had at least 2 mo of experience with the machine and had performed USE on more than 100 nodules during training. The probe was positioned perpendicular to the skin, and repetitive compression was applied above the targeted thyroid nodules during USE. A square region of interest was placed at the target nodule, with the superior margin including subcutaneous fat and the inferior margin including the longus colli muscle. Color homogeneity within the region and a pressure indicator (range: 2–3) were monitored for optimal image acquisition (Moon et al. 2012). In the split-screen mode, gray-scale US images were displayed on the right, and USE images superimposed on the corresponding gray-scale US images were displayed on the left. USE images were displayed with 256 specific colors for each pixel from a color spectrum of red to blue. The softest component was displayed in red and the stiffest component in blue (Moon et al. 2012). USE images were obtained as video files with more than 5 s of continuous length.

Image interpretation

Stored gray-scale images and USE video files were reviewed by one radiology resident (J.E.K.). Appropriate transverse and longitudinal views of each nodule were manually captured after review of gray-scale US images on PACS (picture archiving and communication system). All clinical data were removed from images. Images of each nodule were assigned random numbers and ordered. Video files less than 5 s long were excluded during the USE video file review. USE video files of each nodule were also assigned random numbers different from those of the gray-scale US images.

Three radiologists from three different hospitals retrospectively evaluated the gray-scale US images and USE video files. The first radiologist (H.J.M.) had 11 y of thyroid US experience and 5 y of USE experience. The second radiologist (J.S.P.) had 12 y of thyroid US experience and 8 y of USE experience. The third radiologist (S.J.K.) had 7 y of thyroid US experience and 5 y of USE experience. All three observers were unaware of clinical data or cytologic results. First, gray-scale US images were sent to each observer for evaluation, and interpreted results were recorded in a report collected immediately after image review. Three months after the gray-scale US image review, a set of USE video files were sent to each observer, and the results were then recorded in another report and collected.

On the gray-scale US image interpretation report, five features of the thyroid nodules were recorded. The internal composition of nodules was recorded as solid, $<50\%$ cystic, $\geq 50\%$ cystic and a cyst. Echogenicity of

nodules was recorded as hyper-echogenic, iso-echogenic, hypo-echogenicity and markedly hypo-echogenic. The margin of nodules was evaluated as well circumscribed, microlobulated or irregular. Presence of calcification in nodules was recorded as microcalcification, macrocalcification (including egg shell calcification), mixed micro- and macrocalcification and no calcification. Shape was interpreted as taller than wide or wider than tall. Features of malignant thyroid nodules included solid internal composition, marked hypo-echogenicity, microlobulated or irregular margin, presence of microcalcification and taller than wide shape (Kwak et al. 2011). Final assessment was recorded on the basis of the presence of malignant features, with an assessment of probably benign when there were no malignant features and of suspicious malignant when one or more malignant features were present. There were 37 thyroid nodules containing macrocalcifications, and two thyroid nodules had predominantly cystic portions.

Ultrasound elastography video files were reviewed, and thyroid nodules were scored according to criteria of Asteria et al. (2008) and Rago et al. (2007) separately (Fig. 1). Asteria et al. scored elasticity (Asteria criteria) on a 4-grade scale, with increasing grade denoting decreasing elasticity. Scores of 3 and 4 were classified as suspicious malignant, and scores of 1 and 2 as probably benign. The scoring system used by Rago et al. (Rago criteria) used a 5-grade scale (1–5), with increasing grade denoting decreasing elasticity. Nodules with scores of 4 and 5 were classified as suspicious malignant, whereas scores of 1 to 3 were classified as probably benign.

US-guided FNA

Ultrasound-guided FNA was performed by the same radiologist who performed US. Aspiration was performed at least twice for each nodule using the freehand technique with a 23-gauge needle attached to 2-mL disposable plastic syringe. Obtained samples were expelled onto glass slides, smeared and immediately placed into 95% alcohol for Papanicolaou staining. One of five cytopathologists specializing in thyroid cytology interpreted the smeared samples. The Bethesda classification was used in cytology reports for thyroid aspirate samples (Cibas and Ali 2009).

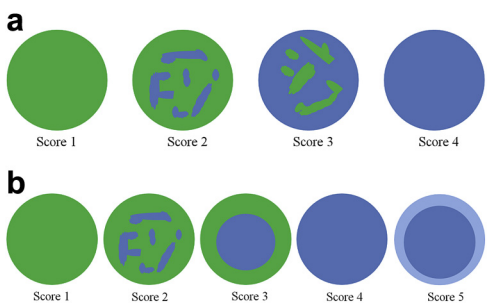

Fig. 1. Schematic representation of elasticity of a thyroid nodule scored according to (a) Asteria criteria and (b) Rago criteria. (a) 1 = elasticity in the whole examined area, 2 = elasticity in a large portion of the examined area, 3 = no elasticity in a large portion of the examined area, 4 = no elasticity in the whole examined area, (b) 1 = elasticity in the whole nodule, 2 = elasticity in a large part of the nodule, 3 = elasticity only in the periphery of the nodule, 4 = no elasticity in the nodule, 5 = no elasticity in the nodule and in the posterior shadowing.

Data and statistical analysis

We used cytopathologic results as the reference standard, with samples confirmed as malignant through pathology or FNA being classified in the positive group, and samples confirmed as benign through pathology or FNA being classified in the negative group. Continuous variables were analyzed with Student's *t*-test, and categorical variables were analyzed with Pearson's χ^2 test. Inter-observer variability of gray-scale US and USE was evaluated between two observers using Cohen's κ analysis for pairwise comparison. For overall comparison among the three observers of gray-scale US and USE, we used the generalized κ with the Inter_Rater SAS Macro program. Additionally, we compared κ coefficients among the final assessments of gray-scale US and elastography according to the Asteria and Rago criteria (Gwet 2002). The relative strength of agreement associated with κ statistics was classified as poor ($\kappa \leq 0.2$), fair ($0.2 < \kappa \leq 0.4$), moderate ($0.4 < \kappa \leq 0.6$), substantial ($0.6 < \kappa \leq 0.8$) or good ($\kappa > 0.8$) (Landis and Koch 1977). We calculated and compared the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and accuracy in predicting malignancy on gray-scale US with the values for the combination of gray-scale US and USE, for each observer, with the generalized estimating equation method. Analysis was performed using SAS Version 9.2 (SAS Institute, Cary, NC, USA.). Statistical significance was assumed when the p value was <0.05.

RESULTS

Of the total 443 nodules, 240 (54.2%) nodules were found malignant on follow-up FNA or surgery, and the remaining 203 (45.8%) nodules were benign. Sixty nodules were diagnosed as malignant on follow-up FNA but did not undergo surgery. Among the 180 nodules surgically confirmed as malignant, 174 were conventional papillary thyroid carcinoma, five were the follicular variant of papillary thyroid carcinoma and one was medullary carcinoma. Among the eight nodules surgically confirmed as benign, four nodules were adenomatous hyperplasia, two were lymphocytic thyroiditis, one was Hurthle cell adenoma and one was cellular adenomatous

hyperplasia. The malignant nodules (9.23 ± 4.26 mm) were smaller than the benign nodules (13.39 ± 6.12 mm), with statistical significance ($p < 0.001$). Gender and age were not associated with malignancy ($p = 0.277$ and 0.074, respectively). Inter-observer agreement was analyzed for each feature in gray-scale US and USE (Table 1). Overall inter-observer agreement was substantial in margin and shape ($\kappa = 0.618$ and 0.760) and moderate in composition, echogenicity and calcification ($\kappa = 0.545$, 0.417 and 0.592). Shape exhibited the highest level of inter-observer agreement, and echogenicity, the lowest level, among the three observers. With respect to the final assessment of gray-scale US, overall inter-observer agreement was substantial ($\kappa = 0.621$) and inter-observer variability between two observers was also substantial among the three observers ($\kappa = 0.603$–0.644).

Ultrasound elastography using the Rago criteria had the lowest overall and pairwise inter-observer agreement compared with gray-scale US or USE using the Asteria criteria ($p < 0.043$) (Fig. 2). Overall inter-observer agreement was substantial ($\kappa = 0.602$) for USE using the Asteria criteria and fair for USE using the Rago criteria ($\kappa = 0.360$) among the three observers. There was substantial pairwise inter-observer agreement between observers 1 and 2 ($\kappa = 0.601$) and between observers 2 and 3 ($\kappa = 0.624$) and moderate agreement between observers 1 and 3 ($\kappa = 0.588$) on USE using the Asteria criteria. On USE using the Rago criteria, inter-observer agreement was moderate between observers 1 and 2 ($\kappa = 0.475$) and between observers 2 and 3 ($\kappa = 0.427$) and fair between observers 1 and 3 ($\kappa = 0.240$). The lowest level of agreement was that between observers 1 and 3 using the Rago criteria.

Diagnostic performance was calculated for gray-scale US, USE and the combination of gray-scale US with USE (Table 2). The sensitivity, NPV and accuracy of gray-scale US (70.4%–80.8%, 68.4%–75.3% and 72.9%–76.5%, respectively) were higher than those of both USE using the Asteria criteria (45.0%–59.2%, 53.2%–57.8% and 58.2%–62.3%, respectively) and USE using the Rago criteria (15.4%–41.3%, 49.0%–53.5% and 52.4%–59.0%, respectively) for all three observers. Sensitivity, NPV and accuracy were lower for USE using the Rago criteria than for USE using the Asteria criteria for all three observers. Specificity was the highest in USE using the Rago criteria (79.8–96.1%) compared with USE using the Asteria criteria (63.6–73.9%) and gray-scale US (69.0–75.9%). PPV was the highest in USE using the Rago criteria (82.2%) compared with USE using the Asteria criteria (67.1%) and gray-scale US (77.5%) for observer 1. Observers 2 and 3 had the highest PPV in gray-scale US (83.0% and 75.5%), compared with USE using the Asteria criteria (64.4% and 67.3%) and USE using the Rago criteria (73.4% and 70.7%).

We compared the diagnostic performances of gray-scale US with that of the combination of gray-scale US with USE (Table 2, Fig. 3). Sensitivity was statistically significantly increased by adding USE using the Asteria criteria to gray-scale US (81.3%–88.3%) and by adding USE using the Rago criteria to gray-scale US (75.4%–85.4%) for all three observers compared with gray-scale US (70.4%–80.8%). NPV was also increased by adding USE using the Asteria criteria to gray-scale US (72.7%, 75.8%) and by adding USE using the Rago criteria to gray-scale US (72.0%, 75.4%) for observers 1 and 2 compared with gray-scale US (68.4%, 70.9%). For observer 3, NPV was also increased by adding USE using the Asteria criteria to gray-scale US (79.0%) and by adding USE using the Rago criteria to gray-scale US (77.4%) compared with gray-scale US (75.3%); however, the differences were not statistically significant ($p = 0.085$ and 0.170). Accuracy was increased by adding USE using the Rago criteria to gray-scale US (74.9%), compared with gray-scale US (72.9%), for observer 1, and for observers 2 and 3, accuracy was decreased by adding USE using the Asteria criteria to gray-scale US (71.6% and 71.6%), compared with gray-scale US (76.5% and 75.4%), with statistical significance. Specificity and PPV were decreased by adding USE using the Asteria criteria to

Table 1. Inter-observer variability of each feature and final assessment of gray-scale US and USE scored according to the Asteria and Rago criteria

	κ, mean (SE)							
	Composition*	Echogenicity*	Margin*	Calcification*	Shape*	Final assessment*	Asteria criteria†	Rago criteria‡
Overall	0.545 (0.044)	0.417 (0.066)	0.618 (0.080)	0.592 (0.045)	0.664 (0.037)	0.621 (0.028)	0.602 (0.028)	0.360 (0.050)
1 *vs.* 2	0.483 (0.046)	0.311 (0.058)	0.618 (0.037)	0.573 (0.046)	0.627 (0.041)	0.620 (0.037)	0.601 (0.037)	0.475 (0.050)
1 *vs.* 3	0.684 (0.038)	0.534 (0.054)	0.636 (0.037)	0.596 (0.045)	0.690 (0.038)	0.644 (0.036)	0.588 (0.038)	0.240 (0.040)
2 *vs.* 3	0.464 (0.045)	0.404 (0.068)	0.603 (0.038)	0.609 (0.046)	0.678 (0.038)	0.603 (0.037)	0.624 (0.037)	0.427 (0.050)

US = ultrasound; USE = ultrasound elastography.
* Gray-scale US.
† Asteria criteria USE scored on a 4-grade scale.
‡ Rago criteria USE scored on a 5-grade scale.

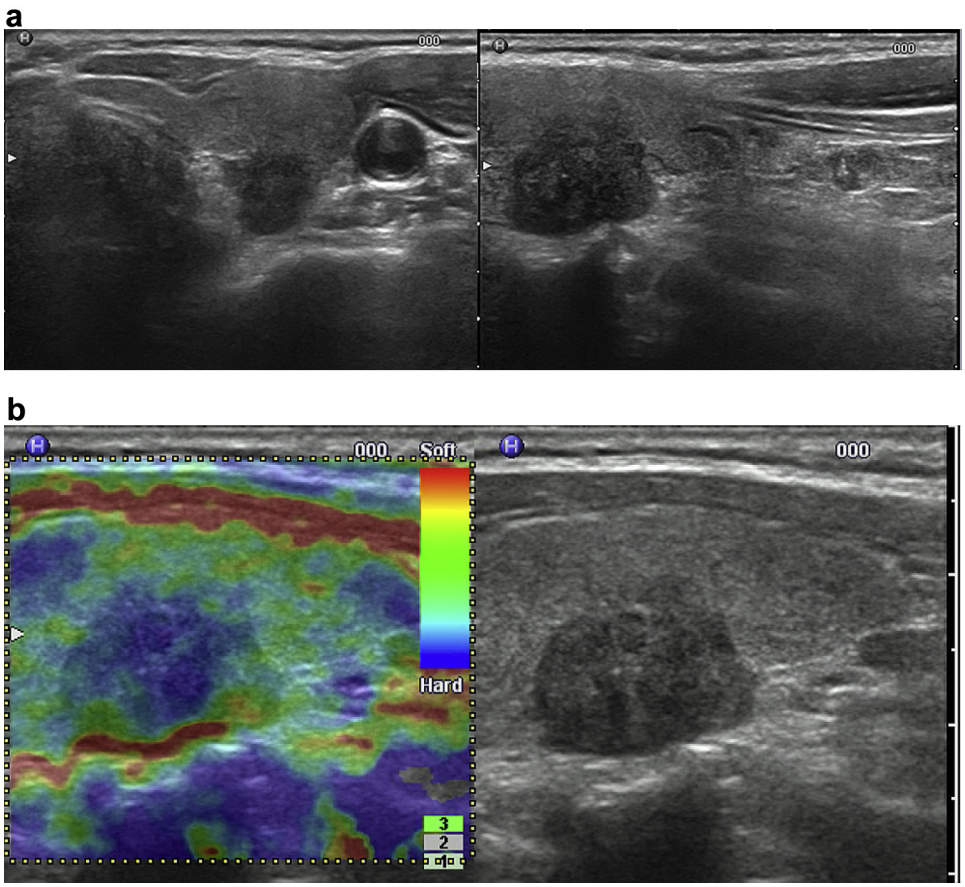

Fig. 2. Representative case of the lowest inter-observer agreement between USE using the Rago criteria and gray-scale US or USE using the Asteria criteria. The patient was a 58-y-old woman with a 15-mm thyroid nodule. (a) All three observers assessed this nodule as suspicious on gray-scale US. (b) On USE, all observers concordantly scored this nodule as 3 using the Asteria criteria, which classified it as suspicious malignant. According to Rago criteria, the observers had discordant scores, with observer 1 and observer 2 scoring it as 2 and 3, benign, and observer 3 scoring it as 4, suspicious malignant. This nodule was surgically confirmed as malignant. USE = ultrasound elastography; US = ultrasound.

gray-scale US (51.7–59.1% and 68.4–70.1%) and by adding USE using the Rago criteria to gray-scale US (59.1%–73.9% and 71.2%–28.3%), compared with gray-scale US (69.0%–82.8% and 75.5%–83.0%), except for observer 1, for whom the addition of USE using the Rago criteria to gray-scale US resulted in no statistically significant difference compared with gray-scale US alone.

DISCUSSION

Ultrasound elastography is a promising technique visualizing the elastic restoring forces of tissue that act against deformation (Bamber et al. 2013). Strain USE uses mechanically induced quasi-static force, and the results of tissue compression and deformation are displayed as an image (Bamber et al. 2013). Malignant nodules tend to be stiffer than benign lesions, and physicians can subjectively differentiate malignant nodules from benign nodules by palpation (Kwak and Kim 2011). On USE, stiff lesions strain less than surrounding soft tissue, and in this manner we can differentiate malignant nodules from benign nodules objectively (Asteria et al. 2008; Azizi et al. 2013; Hong et al. 2009; Kagoya et al. 2010; Mehrotra et al. 2013; Moon et al. 2012; Rago et al. 2007; Rubaltelli et al. 2009; Shuzhen 2012; Shweel and Mansour 2013; Trimboli et al. 2012; Unluturk et al. 2012).

The reproducibility of USE is an important factor in its use as a complementary tool to gray-scale US (Lim et al. 2012; Park et al. 2009; Ragazzoni et al. 2012; Unluturk et al. 2012). Variability can occur in various USE procedures, from selection of imaging planes to compression, selection of images from dynamic sequences and scoring (Lim et al. 2012). There have been reports on factors affecting the poor reliability of USE and the inter-observer and intra-observer agreement for assessing thyroid nodules, including <50% green color in the region of interest box for the thyroid parenchyma, discordance in elasticity scores in the USE images and intra-nodular color signal loss (Kim et al. 2012).

Table 2. Diagnostic performance of gray-scale US and USE scored according to the Asteria and Rago criteria, and addition of USE using the Asteria and Rago criteria to gray-scale US

Observer:	Sensitivity (%)			Specificity (%)			PPV (%)			NPV (%)			Accuracy (%)		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Gray-scale US	70.4	71.3	80.8	75.9	82.8	69.0	77.5	83.0	75.5	68.4	70.9	75.3	72.9	76.5	75.4
USE															
Using Asteria score	45.0	55.8	59.2	73.9	63.6	66.0	67.1	64.4	67.3	53.2	54.9	57.8	58.2	59.4	62.3
Using Rago score	15.4	28.8	41.3	96.1	87.7	79.8	82.2	73.4	70.7	49.0	51.0	53.5	52.4	55.8	58.9
Gray-scale US + USE															
Using Asteria score	81.3 (<0.001)	85.0 (<0.001)	88.3 (<0.001)	59.1 (<0.001)	55.7 (<0.001)	51.7 (<0.001)	70.1 (<0.001)	69.4 (<0.001)	68.4 (<0.001)	72.7 (0.031)	75.8 (0.034)	79.0 (0.085)	71.1 (0.301)	71.6 (0.019)	71.6 (0.019)
Using Rago score	75.4 (<0.001)	79.6 (<0.001)	85.4 (<0.001)	74.4 (0.081)	73.9 (<.001)	59.1 (<0.001)	77.7 (0.817)	78.3 (0.001)	71.2 (<0.001)	71.9 (0.002)	75.4 (0.003)	77.4 (0.170)	74.9 (0.019)	77.0 (0.746)	73.4 (0.105)

PPV = positive predictive value; NPV = negative predictive value; US = ultrasound; USE = ultrasound elastography.

In this study, we evaluated inter-observer variability using video files of USE, excluding technical factors by a performer of USE. Among the features of gray-scale US, shape exhibited the highest level of agreement which was substantial, whereas echogenicity had the lowest level of agreement, which ranged from fair to moderate. These results were found for two observers and also among all three observers. In the final assessment of gray-scale US, there was substantial agreement not only between two observers, but also among the three observers.

For USE using the Asteria criteria, inter-observer variability with gray-scale US features was comparable for two observers as well as among the three observers. Although there was less concordance for USE using the Rago criteria than for gray-scale US, for USE using the Asteria criteria, there was fair to moderate concordance between two observers and fair concordance among the three observers. This difference may result from the different scales; for USE using the Asteria criteria, a 4-grade scoring system was employed, whereas for USE using the Rago criteria, a 5-grade scoring system was employed. In addition, simpler definitions were used with the Asteria criteria than with the Rago criteria. In a previous study, a 4-grade system to score USE similar to that for USE using the Asteria criteria revealed good agreement between two observers and among three observers; these results were similar to ours (Friedrich-Rust et al. 2013; Ragazzoni et al. 2012). Other studies reported excellent inter-observer agreement for USE when using a scoring system different from the system we used and comparing only between two observers (Calvete et al. 2013; Merino et al. 2011).

The reported sensitivity of gray-scale US varies from 83.3% to 94.0%, and specificity varies from 66% to 92.0%, from study to study (Kim et al. 2002; Koike et al. 2001; Moon et al. 2008; Tae et al. 2007). In this study, the sensitivity of USE ranged from 45.0% to 59.1% when the Asteria criteria were used and from 15.4% to 41.3% when the Rago criteria, were used, and specificity ranged from 66.0% to 73.9% with the Asteria criteria and from 79.8% to 96.1% with the Rago criteria, which represent relatively low diagnostic performance compared with initial studies (sensitivity: 94%–97%, specificity: 81%–100%) that used the same scoring system (Asteria et al. 2008; Rago et al. 2007). The diagnostic performance of USE alone was also lower compared with the final assessment of gray-scale US except for specificity in USE using the Rago criteria. However, USE using the Rago criteria exhibited lower sensitivity for each feature in gray-scale US. These results differ from those of previous studies that suggest that high diagnostic performance is possible with USE (Azizi et al. 2013; Hong et al. 2009; Shuzhen 2012; Trimboli et al. 2012).

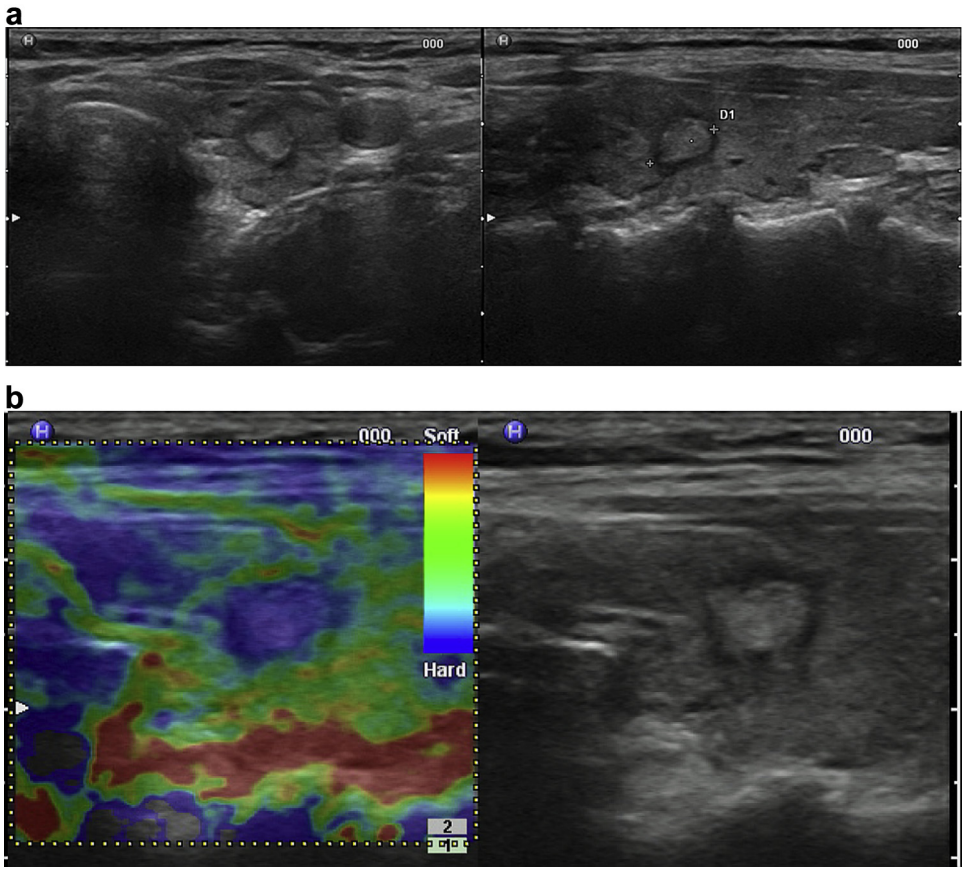

Fig. 3. False-positive case of ultrasound elastography. USE. The patient was a 45-y-old woman with a 14-mm thyroid nodule. (a) On gray-scale US, all three observers assessed this nodule as benign. (b) On USE, all three observers scored this nodule as 4 using Asteria criteria, and observers 1 and 2 both scored it as 4 and observer 3 scored it as 5 using Rago criteria. This nodule was confirmed as benign by fine-needle aspiration. USE = ultrasound elastography; US = ultrasound.

All measures of the diagnostic performance of USE using the Asteria or Rago criteria, except the specificity and PPV of USE using the Rago criteria, were inferior to those of gray-scale US. However, when combined with gray-scale US, USE increased sensitivity and NPV, whereas it decreased specificity and PPV. USE with the Asteria criteria had better sensitivity when added to gray-scale US than USE using the Rago criteria; however, when USE using the Rago criteria was added to gray-scale US, specificity decreased less (Ragazzoni et al. 2012; Trimboli et al. 2012). These findings are comparable to those of other studies that also suggested increased sensitivity and decreased specificity for the combination of USE and gray-scale US (Moon et al. 2012; Trimboli et al. 2012). When USE using the Rago criteria was added to gray-scale US, accuracy statistically significantly increased compared with that of gray-scale US alone for one observer ($p = 0.019$), and for the other two observers, when USE using the Asteria criteria was added to gray-scale US, accuracy decreased compared with that of gray-scale US ($p = 0.019$).

We acknowledge that there are several limitations in this study. First, only eight benign nodules were surgically confirmed; the remaining benign nodules were confirmed by cytologic results, and there may have been false-negative results. Second, we included thyroid nodules with macrocalcifications and cystic portions to evaluate inter-observer variability, which may affect the diagnostic performance of USE. However, as these nodules constituted only a small portion of this study, inter-observer variability caused by macrocalcifications or cystic portions may have had little effect on the results (Bhatia et al. 2011; Rago et al. 2007). Third, although we compared USE using video files, images obtained with this method might be different from those obtained through real-time image evaluation. This study was not a typical inter-observer study designed to perform USE. We reviewed video files obtained by different operators, and because of the retrospective design of this study, there

were limitations on how the environment was controlled during image acquisition. Fourth, radiologists with varying experience performed elastography. This may influence variability during image acquisition. In addition, although we tried to monitor the pressure indicator for adequate pre-compression, strict restrict control was not possible because numerous radiologists performed elastography, and this might have influenced the variability of the study results.

In conclusion, USE using the Asteria criteria has inter-observer variability comparable to that of gray-scale US. However, when USE is added to gray-scale US, the additional diagnostic yield is limited compared with that of gray-scale US alone.

REFERENCES

Asteria C, Giovanardi A, Pizzocaro A, Cozzaglio L, Morabito A, Somalvico F, Zoppo A. US-elastography in the differential diagnosis of benign and malignant thyroid nodules. Thyroid 2008;18:523–531.

Azizi G, Keller J, Lewis M, Puett D, Rivenbark K, Malchoff C. Performance of elastography for the evaluation of thyroid nodules: A prospective study. Thyroid 2013;23:734–740.

Bamber J, Cosgrove D, Dietrich CF, Fromageau J, Bojunga J, Calliada F, Cantisani V, Correas JM, D'Onofrio M, Drakonaki EE, Fink M, Friedrich-Rust M, Gilja OH, Havre RF, Jenssen C, Klauser AS, Ohlinger R, Saftoiu A, Schaefer F, Sporea I, Piscaglia F. EFSUMB guidelines and recommendations on the clinical use of ultrasound elastography: Part 1. Basic principles and technology. Ultraschall Med 2013;34:169–184.

Bhatia KS, Rasalkar DP, Lee YP, Wong KT, King AD, Yuen HY, Ahuja AT. Cystic change in thyroid nodules: A confounding factor for real-time qualitative thyroid ultrasound elastography. Clin Radiol 2011;66:799–807.

Bhatia K, Tong CS, Cho CC, Yuen EH, Lee J, Ahuja AT. Reliability of shear wave ultrasound elastography for neck lesions identified in routine clinical practice. Ultraschall Med 2012;33:463–468.

Calvete AC, Rodriguez JM, de Dios Berna-Mestre J, Rios A, Abellan-Rivero D, Reus M. Interobserver agreement for thyroid elastography: Value of the quality factor. J Ultrasound Med 2013;32:495–504.

Cantisani V, Grazhdani H, Ricci P, Mortele K, Di Segni M, D'Andrea V, Redler A, Di Rocco G, Giacomelli L, Maggini E, Chiesa C, Erturk SM, Sorrenti S, Catalano C, D'Ambrosio F. Q-elastosonography of solid thyroid nodules: Assessment of diagnostic efficacy and interobserver variability in a large patient cohort. Eur Radiol 2014;24:143–150.

Cibas ES, Ali SZ. The Bethesda System For Reporting Thyroid Cytopathology. Am J Clin Pathol 2009;132:658–665.

Fish SA, Langer JE, Mandel SJ. Sonographic imaging of thyroid nodules and cervical lymph nodes. Endocrinol Metab Clin North Am 2008;37:401–417. ix.

Frates MC, Benson CB, Doubilet PM, Kunreuther E, Contreras M, Cibas ES, Orcutt J, Moore FD Jr, Larsen PR, Marqusee E, Alexander EK. Prevalence and distribution of carcinoma in patients with solitary and multiple thyroid nodules on sonography. J Clin Endocrinol Metab 2006;91:3411–3417.

Friedrich-Rust M, Meyer G, Dauth N, Berner C, Bogdanou D, Herrmann E, Zeuzem S, Bojunga J. Interobserver agreement of Thyroid Imaging Reporting and Data System (TIRADS) and strain elastography for the assessment of thyroid nodules. PLoS One 2013;8:e77927.

Gwet K. Computing inter-rater reliability with the SAS system. Stat Methods Inter-rater Reliability Assess 2002;3:1–16.

Hong Y, Liu X, Li Z, Zhang X, Chen M, Luo Z. Real-time ultrasound elastography in the differential diagnosis of benign and malignant thyroid nodules. J Ultrasound Med 2009;28:861–867.

Kagoya R, Monobe H, Tojima H. Utility of elastography for differential diagnosis of benign and malignant thyroid nodules. Otolaryngol Head Neck Surg 2010;143:230–234.

Kim EK, Park CS, Chung WY, Oh KK, Kim DI, Lee JT, Yoo HS. New sonographic criteria for recommending fine-needle aspiration biopsy of nonpalpable solid nodules of the thyroid. AJR Am J Roentgenol 2002;178:687–691.

Kim KA, Kim MJ, Jeon HM, Kim KS, Choi JS, Ahn SH, Cha SJ, Chung YE. Prediction of microvascular invasion of hepatocellular carcinoma: Usefulness of peritumoral hypointensity seen on gadoxetate disodium-enhanced hepatobiliary phase images. J Magn Reson Imaging 2012;35:629–634.

Kim I, Kim EK, Yoon JH, Han KH, Son EJ, Moon HJ, Kwak JY. Diagnostic role of conventional ultrasonography and shearwave elastography in asymptomatic patients with diffuse thyroid disease: Initial experience with 57 patients. Yonsei Med J 2014;55:247–253.

Ko SY, Kim EK, Sung JM, Moon HJ, Kwak JY. Diagnostic performance of ultrasound and ultrasound elastography with respect to physician experience. Ultrasound Med Biol 2014;40:854–863.

Koike E, Noguchi S, Yamashita H, Murakami T, Ohshima A, Kawamoto H. Ultrasonographic characteristics of thyroid nodules: Prediction of malignancy. Arch Surg 2001;136:334–337.

Kovacevic DO, Skurla MS. Sonographic diagnosis of thyroid nodules: Correlation with the results of sonographically guided fine-needle aspiration biopsy. J Clin Ultrasound 2007;35:63–67.

Kwak JY, Han KH, Yoon JH, Moon HJ, Son EJ, Park SH, Jung HK, Choi JS, Kim BM, Kim EK. Thyroid imaging reporting and data system for US features of nodules: A step in establishing better stratification of cancer risk. Radiology 2011;260:892–899.

Kwak JY, Kim EK. Diagnostic performance of quantitative shear wave ultrasound elastography for thyroid cancer. J Korean Thyroid Assoc 2011;4:109–113.

Kwak JY, Kim EK. Ultrasound elastography for thyroid nodules: Recent advances. Ultrasonography 2014;33:75–82.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.

Lim DJ, Luo S, Kim MH, Ko SH, Kim Y. Interobserver agreement and intraobserver reproducibility in thyroid ultrasound elastography. AJR Am J Roentgenol 2012;198:896–901.

Mehrotra P, McQueen A, Kolla S, Johnson SJ, Richardson DL. Does elastography reduce the need for thyroid FNAs? Clin Endocrinol 2013;78:942–949.

Merino S, Arrazola J, Cardenas A, Mendoza M, De Miguel P, Fernandez C, Ganado T. Utility and interobserver agreement of ultrasound elastography in the detection of malignant thyroid nodules in clinical care. AJNR Am J Neuroradiol 2011;32:2142–2148.

Moon WJ, Jung SL, Lee JH, Na DG, Baek JH, Lee YH, Kim J, Kim HS, Byun JS, Lee DH. Benign and malignant thyroid nodules: US differentiation—Multicenter retrospective study. Radiology 2008;247:762–770.

Moon HJ, Sung JM, Kim EK, Yoon JH, Youk JH, Kwak JY. Diagnostic performance of gray-scale US and elastography in solid thyroid nodules. Radiology 2012;262:1002–1013.

Park SH, Kim SJ, Kim EK, Kim MJ, Son EJ, Kwak JY. Interobserver agreement in assessing the sonographic and elastographic features of malignant thyroid nodules. AJR Am J Roentgenol 2009;193:W416–W423.

Ragazzoni F, Deandrea M, Mormile A, Ramunni MJ, Garino F, Magliona G, Motta M, Torchio B, Garberoglio R, Limone P. High diagnostic accuracy and interobserver reliability of real-time elastography in the evaluation of thyroid nodules. Ultrasound Med Biol 2012;38:1154–1162.

Rago T, Santini F, Scutari M, Pinchera A, Vitti P. Elastography: New developments in ultrasound for predicting malignancy in thyroid nodules. J Clin Endocrinol Metab 2007;92:2917–2922.

Rubaltelli L, Corradin S, Dorigo A, Stabilito M, Tregnaghi A, Borsato S, Stramare R. Differential diagnosis of benign and malignant thyroid nodules at elastosonography. Ultraschall Med 2009;30:175–179.

Shuzhen C. Comparison analysis between conventional ultrasonography and ultrasound elastography of thyroid nodules. Eur J Radiol 2012; 81:1806–1811.

Shweel M, Mansour E. Diagnostic performance of combined elastosonography scoring and high-resolution ultrasonography for the differentiation of benign and malignant thyroid nodules. Eur J Radiol 2013;82:995–1001.

Tae HJ, Lim DJ, Baek KH, Park WC, Lee YS, Choi JE, Lee JM, Kang MI, Cha BY, Son HY, Lee KW, Kang SK. Diagnostic value of ultrasonography to distinguish between benign and malignant lesions in the management of thyroid nodules. Thyroid 2007;17: 461–466.

Takashima S, Fukuda H, Nomura N, Kishimoto H, Kim T, Kobayashi T. Thyroid nodules: Re-evaluation with ultrasound. J Clin Ultrasound 1995;23:179–184.

Trimboli P, Guglielmi R, Monti S, Misischi I, Graziano F, Nasrollah N, Amendola S, Morgante SN, Deiana MG, Valabrega S, Toscano V, Papini E. Ultrasound sensitivity for thyroid malignancy is increased by real-time elastography: A prospective multicenter study. J Clin Endocrinol Metab 2012;97:4524–4530.

Unluturk U, Erdogan MF, Demir O, Gullu S, Baskal N. Ultrasound elastography is not superior to grayscale ultrasound in predicting malignancy in thyroid nodules. Thyroid 2012;22:1031–1038.

Veyrieres JB, Albarel F, Lombard JV, Berbis J, Sebag F, Oliver C, Petit P. A threshold value in shear wave elastography to rule out malignant thyroid nodules: A reality? Eur J Radiol 2012;81:3965–3972.

Yoon JH, Yoo J, Kim EK, Moon HJ, Lee HS, Seo JY, Park HY, Park WJ, Kwak JY. Real-time elastography in the evaluation of diffuse thyroid disease: A study based on elastography histogram parameters. Ultrasound Med Biol 2014;40:2012–2019.

Zhang YF, Xu HX, He Y, Liu C, Guo LH, Liu LN, Xu JM. Virtual Touch tissue quantification of acoustic radiation force impulse: A new ultrasound elastic imaging in the diagnosis of thyroid nodules. PLoS One 2012;7:e49094.